INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Why do ASR systems despite neural nets still depend on robust features

Angel Mario Castro Martinez, *Marc René Schädler*

Medizinische Physik and Cluster of Excellence Hearing4all
Carl von Ossietzky Universität Oldenburg, Germany
{angel.castro, marc.r.schaedler}@uni-oldenburg.de

Abstract

To which extent can neural nets learn traditional signal processing stages of current robust ASR front-ends? Will neural nets replace the classical, often auditory-inspired feature extraction in the near future? To answer these questions, a DNN-based ASR system was trained and tested on the Aurora4 robust ASR task using various (intermediate) processing stages. Additionally, the training set was divided into several fractions to reveal the amount of data needed to account for a missing processing step on the input signal or prior knowledge about the auditory system. The DNN system was able to learn from ordinary spectrograms representations outperforming MFCC using 75% of the training set and almost as good as log-Mel-spectrograms with the full set; on the other hand, it was unable to compensate the robustness of auditory-based Gabor features, which even using 40% of the training data outperformed every other representation. The study concludes that even with deep learning approaches, current ASR systems still benefit from a suitable feature extraction.

Index Terms: automatic speech recognition, deep neural networks, resource constrained training, robust front-ends

1. Introduction

Recently, neural nets saw their (second) renaissance as state-of-the-art recognizers and are now ubiquitous in automatic speech recognition (ASR) systems. As they are capable of learning high-dimensional non-linear functions [1, 2], they have partially, if not completely, taken over the feature extraction module, blurring the line between front-end and back-end. Hence, it is a legitimate question if traditional (robust) ASR front-ends which implement, to some extent, prior knowledge about the human auditory system, will be replaced by neural nets.

Ever since their introduction Mel-frequency cepstral coefficients (MFCC) have been broadly used in ASR systems, owing to their few dimensions and relatively easy procedure [3]. Lately, however, more primitive representations have been adopted mainly as a consequence of the implementation of deep neural networks (DNN); for instance logarithmic scaled Mel-spectrogram from which MFCC are extracted, dubbed as Mel-frequency spectral coefficients (MFSC).

There have been various studies approaching this subject from a different view [4, 5, 6, 7], searching for models trained on simpler features to outperform the MFSC and/or their variants, going even to the point of training acoustic models with raw discrete signals. In [8], using a combination of convolutional, recurrent and fully connected neural networks this endeavor was finally achieved.

On the other hand, this last model was trained with over 2000 hours of speech material. A crucial prerequisite for training neural nets is the availability of sufficient training data as performance directly depends on it; hence, reliable labeled data may be one of the most valuable resources.

If there were common fundamental principles of signal processing for robust speech recognition and we knew how to implement them, it would probably be more beneficial to use the available training data to learn *unknown* signal properties instead of *known* signal processing principles. Learning static components or —"re-inventing the wheel"— every single time an ASR system is trained could be a waste of valuable training resources, if an adequate implementation of these components existed.

In this study we assessed whether a current robust front-end together with a DNN-based recognizer could be such an adequate implementation or if it can be replaced by preceding processing stages. On a widely available robust ASR task, namely the Aurora 4 framework, the amount of training data was systematically varied from 10% to 100%, which approximately corresponds to 1.5 to 15 hours, respectively.

The robust front-end proposed in [9] used a Gabor filter bank (GBFB [10]) of auditory-inspired spectro-temporal modulation filters to extract spectro-temporal patterns and was reported to outperform several current robust ASR front-ends.

It implements basic human auditory signal processing principles, such as the sound intensity compression and the pitch perception of pure tones, in addition, it encodes spectro-temporal changes in the signal using a filter bank of modulation filters which were inspired from physiological measurements in mammals [11, 12].

If representations as such are suitable for robust ASR, their use could prevent the ASR system from learning common signal processing principles and allow to "invest" the training data in learning the unknown signal properties; which, in turn, should result in increased recognition performance or allow to achieve the same performance with less training data. Conversely, if a stage of the fixed signal processing is sub-optimal for robust ASR, the recognition performance should decrease over using the preceding stage.

To assess to which extent the neural nets can learn or even improve missing stages of the feature extraction using a DNN-based ASR system, several intermediate stages of two front-ends were considered: The traditionally used MFCC features and robust auditory-inspired Gabor filter bank features. Both share the same intermediate stage, the logarithmic-scaled Mel-spectrogram, which, in turn, is calculated from an ordinary amplitude spectrogram, these four processing steps were considered as features to train individual models.

http://dx.doi.org/10.21437/Interspeech.2016-1552

2. Methods

2.1. Recognition experiments

Experiments were performed on the large vocabulary continuous speech recognition Aurora 4 framework [13], derived from the LDC Wall Street Journal Corpus. The multicondition train set consists of 7137 utterances from 83 independent speakers (~15hrs.); half of them were recorded using a close-talk microphone and the other half a secondary microphone to introduce some channel distortions; each half was further divided in 7 groups leaving the first one unprocessed and on the remaining 6, an individual additive noise was introduced (car, babble, restaurant, street, airport and train) at random signal-to-noise-ratios from (10 - 20 dB). The 5000-word-vocabulary test set includes 330 utterances from 8 independent speakers; each utterance was processed following the 14 aforementioned conditions (except the SNR levels of the additive noise ranged 5 - 15 dB) resulting into 4620 utterances.

2.2. Robust front-end

In this section we describe the features as a series of processing steps. Starting with the amplitude spectrogram (Raw-Spec) obtained by applying a 512 point FFT to the frame segmented (25 ms length and 10 ms shift) discrete signal and keeping the absolute values of the first 257 coefficients. In the second step, a Mel-spectrogram (Mel-spec) is calculated by combining the 257 frequency bins into 31 equidistant Mel-bands using triangular filters. This filtering mimics auditory signal processing principle in that it limits the spectral resolution and represents frequencies similar to their distribution on the human basilar membrane. The third step is to apply a logarithmic compression to the amplitude values encoded in the Mel-spectrogram to obtain a log-Mel-spectrogram, which mimics to some extent the compression of sound intensity in human auditory perception. These Mel-frequency spectral coefficients (MFSC) are the precursors of the last two diverging processing steps we compared: MFCC (plus deltas and double deltas) and Gabor filterbank (GBFB) features.

Gabor features derive from the convolution between the MFSC and a set of 2-dimensional Gabor filters introduced by [10]. A Gabor filter is the product of a complex sinusoid function (1) and a Hann window (2). The periodicity of the carrier sinusoid was defined by the radian frequencies ω_n and ω_k (n and k denoting time and frequency index, respectively), in this study we kept the temporal modulation frequency fixed at 25 Hz (referred to as high temporal modulation or HTM) to limit the temporal context to ± 3 frames while varying ω_k from -0.25 to $+0.25$ cycles/channel, allowing the Gabor filters to be tuned to particular spectro-temporal directions. These Gabor features are dubbed GBFB-HTM.

$$s(n,k) = \exp\big(i\omega_n(n-n_0) + \omega_k(k-k_0)\big) \tag{1}$$

$$h(n,k) = \frac{1}{2} - \frac{1}{2}\cos\left(\frac{2\pi(n-n_0)}{W_n+1}\right)\cos\left(\frac{2\pi(k-k_0)}{W_k+1}\right) \tag{2}$$

Because the output of the filtering process with filters of low spectral modulation frequencies and hence large spectral extent (69 channels height for $\omega_k = 0$) is highly redundant in the spectral dimension, it is critically sampled in the spectral dimension to reduce redundancy of the feature dimensions. This step is explained in detail in [10].

2.3. DNN Recognizer

The deep neural network (DNN) was based on the one described in [14]. Recognition experiments were conducted using the Kaldi ASR toolkit [15]. Due to the size restrictions we imposed on the train set, we pretrained the acoustic models using a stack of Restricted Boltzmann Machines [16], also known as deep belief network (DBN), as it provides a substantial improvement on low resource training data over random initialization [17].

The 7-layers DBN served as a backbone for the final network, it was later fine-tuned using back-propagation via SGD to classify frames into triphone-states using an independent (development) set and the cross entropy between the network output and the labels as a cost function. A regular GMM triphone system was trained in order to obtain those labels via forced alignment.

The training was done in up to 20 epochs (stopping when the relative improvement was lower than 0.001). The starting learning rate was 0.008 (halving it every time the relative improvement was lower than 0.01) and no momentum nor regularization techniques were applied. Each of the six hidden layers in the resulting DNN has 2048 sigmoid neurons.

To provide temporal context several frames were concatenated and parsed to the input layer, ± 1 frames for GBFB-HTM and ± 5 frames for the remaining features. A soft-max layer of approximately 2000 units was attached to the end of the DNN to output the most likely posterior probabilities of each context-dependent HMM state.

2.4. Training data limitation

To keep track of the amount of data needed for each model to account for missing processing steps, the train set was divided first in halves and quarters, then percentagewise to add additional values in between; producing the following steps: 10%, 25%, 40%, 50%, 60%, 75%, 90% and 100%. Every percentual portion includes all 83 speakers and a distributed randomized selection of utterances per speaker, to ensure there was a balanced number of examples per target class (i.e. triphones) for each condition; however, every minor portion was a subset of a bigger one. The random selection was performed 10 times to get a margin of error on every percentage step. The test set remained the same for every model.

3. Results

The word error rates (WERs) on the Aurora 4 task depending on the employed front-end and available training data are plotted in Figure 1 and reported in numerical form in Table 1.

The WERs of all systems increased monotonically as the available training data was decreased. As expected, the effect was more pronounced when fewer samples were available, e.g. from 25% to 10%. Below 10% no reliable training of the system was possible anymore.

By limiting the training data to 10% the WERs almost doubled compared to models taking advantage of the whole training data set, regardless of the front-end. Independent of the amount of training data, using the GBFB-HTM features resulted in 4-6 percentage points lower WERs than the lowest achieved WERs with the other front-ends, outperforming these by far. Hence, the best performance, i.e., the lowest WER of (10.05 ± 0.05)%, was achieved when using GBFB-HTM features and the whole training data set.

Surprisingly, using the traditional signal representation offered by the MFCC did not result in second lowest WERs. With

Table 1: *Word error rates in % depending on the available training data and front-end on the Aurora 4 task along with the uncertainty due to the selected portion of the training data and initial values for the DNN estimated from 10 runs. 100% training data corresponds to about 15 hours of speech recordings.*

Front-end	10%	25%	40%	50%	60%	75%	90%	100%
1) Raw-Spec	27.40 ±0.29	21.38 ±0.10	19.35 ±0.05	17.61 ±0.19	16.71 ±0.08	15.42 ±0.09	14.86 ±0.11	14.13 ±0.06
2) Mel-Spec	30.50 ±0.37	21.00 ±0.24	19.36 ±0.13	18.58 ±0.15	16.61 ±0.12	16.34 ±0.24	15.60 ±0.12	14.97 ±0.20
3) MFSC	25.49 ±0.27	21.61 ±0.09	17.27 ±0.09	16.90 ±0.05	15.16 ±0.07	14.26 ±0.08	13.82 ±0.10	13.58 ±0.16
4a) MFCC	26.32 ±0.13	20.58 ±0.15	18.46 ±0.16	17.65 ±0.21	16.23 ±0.21	15.93 ±0.16	15.73 ±0.25	14.94 ±0.15
4b) GBFB-HTM	19.84 ±0.27	14.20 ±0.06	12.86 ±0.05	12.26 ±0.06	11.56 ±0.05	10.95 ±0.05	10.34 ±0.06	10.10 ±0.05

the exception of when training on 25% of the training data, where all non-GBFB front-ends performed similarly (± 1 percentage point), using the MFSC as the front-end resulted in the second lowest WER.

While the extraction of GBFB-HTM features from the MFSC greatly reduced the WERs, the extraction of MFCC from the very same representations even increased the WERs (with a small exception when training on 25% of the training data).

Neither preceding processing stages of the MFSC, i.e., the Mel-spectrogram nor the raw amplitude spectrogram, achieved WERs as low with the MFSC themselves, when 40% or more of the training data are used.

Using the raw amplitude spectrogram tended to result in slightly lower WERs compared to when using the Mel-spectrogram. When using 75% or more of the training data, it even outperformed MFCC with the respective WERs being about 0.8 percentage points lower.

The system with GBFB-HTM features trained on only 25% of the training data outperformed the system using MFCC features trained on the full training data set and with only 40% of the training data, it also outperformed any other system using an intermediate signal representation trained even on the full training data set.

Unlike representations obtained from MFCC or their intermediate processing steps, the ones trained from GBFB-HTM yielded steadily decreasing WER when adding more training data.

Furthermore, the uncertainty due to the used portion of the training data and the initialization of the parameters of the neural net as reported in Table 1 when using GBFB-HTM feature was comparatively low, except when using only 10% of the training data.

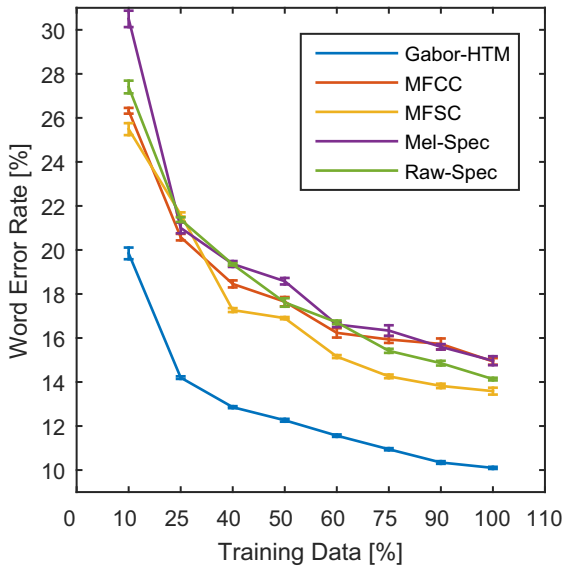

Figure 1: *Word error rates in % depending on the available training data and front-end on the Aurora 4 task. The common intermediate processing stage of 4a) MFCC and 4b) GBFB-HTM features is 3) a logarithmic scaled Mel-spectrogram which conforms the MFSC and in turn are calculated by compressing the amplitude values of 2) a Mel-spectrogram (Mel-Spec) which, in turn, is derived from 1) an amplitude spectrogram (Raw-Spec) by spectral integration.*

4. Discussion

As shown in our previous study [9] the robust GBFB-HTM front-end could not be omitted without important increases in WERs in a DNN-based ASR system performing a current ASR task. The results in this study, suggest a robust front-end, here in the form of GBFB-HTM features, still provides an important advantage compared to learning feature representations from intermediate signal representations even when restricting the amount of available training data.

Compared to their shared intermediate representation—the MFSC— the auditory-inspired GBFB-HTM features aided the ASR system in performing the task while the traditionally used MFCC even harmed the recognition performance. This indicates, that the signal representation by MFCC was sub-optimal for this robust ASR task, because the DNN could learn a more suitable representation from the MFSC, or even from the raw amplitude spectrogram.

In contrast, the signal representation learned using GBFB-HTM features was the most suitable for robust ASR because the DNN could not learn a more suitable representation from neither the MFSC, nor any preceding intermediate representation.

The extraction of spectro-temporal patterns from MFSC could be remarked as a more efficient usage of the available training data based on the consistent WER decrease. Compared to the MFSC, the direct encoding of spectro-temporal patterns could save more than half of the training data and would still achieve lower WERs ((12.26 ± 0.06) vs. (13.58 ± 0.16)).

The GBFB-HTM front-end is the one that implements the most (prior) knowledge about the auditory system, namely, in addition to the processing of MFSC, it considers spectro-

temporal modulation patterns. This extra information might explain the gap between processing steps.

The human auditory system is known to show superior speech recognition performance under the most adverse acoustic conditions and it could well be, that a set of common signal processing principles exists which is near-optimal for robust speech recognition (and potentially many other tasks). The results in this study suggest that such a representation might exist and it might reassemble more GBFB-HTM features than MFCC features.

The inclusion of even more training data (more than 100%) could result in further improvements for any of the ASR systems, but the results give no hint about the lowest achievable WER given additional training data. Hence, there is no point in speculating about the best front-end if (much) more training data was available. Moreover, the performance would most probably depend on the type of additional training data, such as, more speakers, more vocabulary, more noise conditions, etc ...

The availability of more comprehensive training data sets could possibly help to find signal representations which are even better suited for (robust) ASR.

In theory, one could expect an improvement in performance with every adequately implemented bit of the hypothesized common signal processing principles, however, it was not the case for the studied processing chain. Compared to 1) the raw spectrogram, using 2) the Mel-spectrogram as the front-end often resulted in higher WERs, which indicates this processing step might not be adequately implemented and could be possibly improved. Potential alternatives can be found in many different models of human auditory signal processing, e.g., [18].

We hypothesize that, as long as the limited availability of training data is a concern, robust front-end will remain an important part of robust ASR systems, because they alleviate the demand for training data by providing prior knowledge about a suitable signal representation.

Neural nets are an interesting and powerful tool to assess the benefit of specific processing stages by reducing the available training data.

4.1. Future work

The proposed method can be used to examine and reconsider the signal processing stages of current robust ASR front-ends. Certainly, neural nets should be used to evaluate the benefit and reconsider the indispensability of certain signal processing principles (for robust ASR).

It might even be helpful to build new robust front-end from scratch by evaluating the most suitable options step-by-step. Therefore, it could be worthwhile to study the interaction of front-ends and training data if much more, and more comprehensive, training data become available. It should be tested if the findings of this study translate to other robust ASR tasks and other languages in which finding suitable labeled data might be a major shortcoming for using modern recognizers.

5. Conclusions

The most important findings of this work can be summarized as follows:

- Among the considered front-ends, the most robust one was found to be the one that implemented the most auditory-inspired signal processing principles: The Gabor filter bank (GBFB)-based features which encode complex spectro-temporal patterns outperformed all other front-ends even with as much as 60% less training data available.
- Current ASR systems with complex deep learning architectures and training methods will probably benefit from robust features, or implementing a-priori knowledge about suitable signal processing, in challenging acoustic scenes, because the available training data can be more efficiently used to learn unknown signal properties instead of known (auditory) signal processing principles.
- Neural nets can be used as a tool to test if a given front-end, i.e., a fixed signal representation, is suitable for a given task, and if so, to estimate the equivalent of training data the implementation of prior knowledge saved.

6. Acknowledgements

This work was funded by the Cluster of Excellence Hearing4All.

7. References

[1] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems (MCSS)*, vol. 5, no. 4, pp. 455–455, 1992.

[2] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[3] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[4] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5884–5887.

[5] D. Palaz, R. Collobert, and M. M. Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," *arXiv preprint arXiv:1304.1018*, 2013.

[6] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for LVCSR." in *INTERSPEECH*, 2014, pp. 890–894.

[7] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4624–4628.

[8] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," in *Proc. Interspeech*, 2015.

[9] A. Castro Martinez, N. Moritz, and B. T. Meyer, "Should deep neural nets have ears? the role of auditory features in deep learning approaches," in *Proc. INTERSPEECH*, September 2014, pp. 2435–2439.

[10] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 4134–4151, 2012a.

[11] A. Qiu, C. E. Schreiner, and M. A. Escabi, "Gabor analysis of auditory midbrain receptive fields: spectro-temporal and binaural composition," *Journal of Neurophysiology*, vol. 90, no. 1, pp. 456–476, 2003.

[12] N. Mesgarani, D. Stephen, and S. Shamma, "Representation of phonemes in primary auditory cortex: how the brain analyzes speech," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE*, 2007.

[13] N. Parihar, J. Picone, D. Pearce, and H.-G. Hirsch, "Performance analysis of the aurora large vocabulary baseline system," in *Signal Processing Conference, 2004 12th European*, 2004, pp. 553–556.

[14] K. Vesel, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," *Proc. Interspeech*, 2013.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, and N. G. und K. Vesel, "The kaldi speech recognition toolkit," *Proc. ASRU*, 2011.

[16] G. Hinton, "A practical guide to training restricted boltzmann machines," *Momentum*, vol. 9, no. 1, p. 926, 2010.

[17] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *The Journal of Machine Learning Research*, vol. 11, pp. 625–660, 2010.

[18] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 433–442, 2002.